An Exploration in Textual Analysis

Daniels, Grady

The internet's expansion has increased the data available to researchers, allowing analysis to make insights into human behavior. This project statistically analyzes trends in computer science by studying the ArXiv archive's CS articles between 1989 and 2016. Analysis show that documents' meaningful words have a quantitative significance that a machine can extract without understanding the text. Measurements were taken from a sample of ~8,600 documents. Their words were positively weighted by their normalized frequency and negatively weighted by either their presence in other documents (TF-IDF) or by their probability to appear in the body of a web page, in two separate measurements, to reduce the noise in plain word frequencies from "structure words" like "the" and "a." The observed probabilities of words with large weights were graphed to show their popularity during each year. Finally, vector representations of the words were used to find contextually similar words and to examine how changes in popularity may be caused by shifts in terminology. The secondary word relevance measurement was found comparable to TF-IDF, an algorithm used by search engines to retrieve documents: 54.9% of the 30 highest weighted words (secondary measure) are also in the 30 highest TFIDF weighted words. A significant design challenge was testing the integrity of measurements as there is little control data; this was solved by a secondary dataset of words' probability to appear in web pages. In both measurements, words with large weights are hand-verifiably "important" to documents that they appear in, so the results show that words which are meaningful to documents have an identifiable quantitative significance.

Awards Won:

Mu Alpha Theta, National High School and Two-Year College Mathematics Honor Society: Second Award of $1,500